Computational and Mathematical Organization Theory manuscript No.
(will be inserted by the editor)

On the robustness of centrality measures against link weight quantization in social networks

Sho Tsugawa · Yukihiro Matsumoto · Hiroyuki Ohsaki

Received: date / Accepted: date

Abstract In social network analysis, individuals are represented as nodes in a graph, social ties among them are represented as links, and the strength of the social ties can be expressed as link weights. However, in social network analyses where the strength of a social tie is expressed as a link weight, the link weight may be quantized to take only a few discrete values. In this paper, expressing a continuous value of social tie strength as a few discrete value is referred to as *link weight quantization*, and we study the effects of link weight quantization on centrality measures through simulations and experiments utilizing network generation models that generate synthetic social networks and real social network datasets. Our results show that (1) the effects of link weight quantization on the centrality measures are not significant when determining the most important node in a graph, (2) conversely, a 5–8 quantization level is needed to determine other important nodes, and (3) graphs with a highly skewed degree distribution or with a high correlation between node degree and link weights are robust against link weight quantization.

Keywords Social Network · Centrality · Link Weight Quantization · Robustness

1 Introduction

Research on social network analysis has been actively pursued (Watts, 2007; Borgatti et al, 2009). In social network analysis, individuals are represented as nodes in a graph and social

An earlier version of this paper was presented at "The 4th Annual Workshop on Simplifying Complex Networks for Practitioners (SIMPLEX 2012)" in Lyon in 2012. We extended the workshop paper including extensive experiments and discussions.

S. Tsugawa ✉
Faculty of Engineering, Information and Systems, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan
Tel.: +81-29-853-5382
E-mail: s-tugawa@cs.tsukuba.ac.jp

Y. Matsumoto
Graduate School of Information Science and Technology, Osaka University

H. Ohsaki
School of Science and Technology, Kwansei Gakuin University

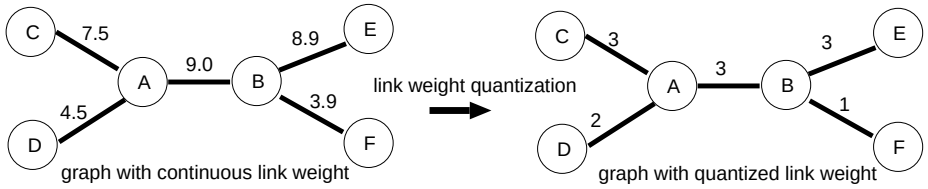

Fig. 1: An example of link weight quantization: The left graph represents a social network, and the strength of a social tie is expressed as a continuous value of link weight. The right graph also represents a social network, but the strength of a social tie is expressed as a few discrete value of link weight. Since a traditional social network analysis uses questionnaires to obtain the tie strength, the right graph is used instead of the left graph for the analysis. In this paper, expressing a continuous value of social tie strength as a few discrete value is referred to as *link weight quantization*.

ties among them, such as similarities, social relations, interactions, and flows, are represented as links (Watts, 2007; Borgatti et al, 2009). The strength of the social tie can be expressed as a link weight. The resulting graph is then analyzed to provide an understanding of complex social phenomena that involve interactions among a large number of people.

Among various indices proposed for social network analysis, centrality measures (degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality) (Freeman, 1979; Bonacich, 1972) have been widely used in actual analyses (Borgatti, 2006; Batallas and Yassine, 2006). Centrality measures are indices that express the influence of one node on others, and have been used for several purposes, such as discovering a person who plays a central role in a community (Borgatti, 2006; Batallas and Yassine, 2006), or inferring activity and leadership levels in a community (Kamei et al, 2008; Tsugawa et al, 2012).

Recently, the use of link weights to express the strength of social ties has become popular, particularly among researchers studying social networks obtained from communication logs (Ehrlich and Cataldo, 2012; Cataldo and Ehrlich, 2012; Gómez et al, 2008), and it has been suggested that using link weights is an effective tool for social network analysis (Opsahl et al, 2010; Opsahl and Panzarasa, 2009). By definition, the strength of a social tie is a "combination of the amount of time, the emotional intensity, the intimacy (mutual confiding) and reciprocal services which characterize the tie" (Granovetter, 1973). Thus, a link weight representing the tie strength should be continuous.

However, in many traditional social network analyses (Batallas and Yassine, 2006; Costenbader and Valente, 2003; Valente et al, 1997), only the existence of a social tie is used and its strength is ignored. Even in social network analyses where the strength of a social tie is expressed as a link weight, the link weight may be quantized to take only a few discrete values (Creswick and Westbrook, 2010; Cross and Parker, 2004) since the strength of a social tie is generally known from the results of questionnaires given to the participants in the experiments. In this paper, expressing a continuous value of social tie strength as a few discrete value is referred to as *link weight quantization*. Figure 1 shows an example of link weight quantization. Link weight quantization should affect the results of social network analyses.

Several analyses on the robustness of centrality measures used for social network analyses against the imperfections of graphs (i.e., noise due to random addition and deletion of nodes and links) have been performed (Borgatti et al, 2006; Lee et al, 2006; L.Frantz et al, 2009; Kim and Jeong, 2007; Costenbader and Valente, 2003). However, since unweighted

graphs are used for the analyses in those studies, effects of ignoring link weight and link weight quantization on centrality measures have not been explored.

Understanding the effects of link weight quantization on the results of social network analysis is important. For instance, such understanding would help social network researchers to know whether using a graph with quantized link weights is enough, or whether using continuous values of link weights is necessary to answer their research questions.

In this paper, we study, through simulations utilizing network generation models that generate synthetic social networks, the effects of link weight quantization on the popular centrality measures (degree, betweenness, closeness, and eigenvector centralities). Following (Borgatti et al, 2006; L.Frantz et al, 2009), we perform several simulations to investigate how centrality measures change when link weights are quantized to take varying numbers of discrete values. We also perform experiments with real social networks to validate our simulation results.

The remainder of this paper is organized as follows. Section 2 introduces related work. In Section 3, the experimental methods are explained. In Sections 4 and 5, we present the experimental results from network generation models and real social networks, and discuss the effects of ignoring link weight and link weight quantization on the centrality measures. Finally, Section 6 contains our conclusions and a discussion of future work.

2 Related Work

Several analyses on the robustness of centrality measures used for social network analyses against the imperfections of graphs have been performed (Borgatti et al, 2006; Lee et al, 2006; L.Frantz et al, 2009; Kim and Jeong, 2007; Costenbader and Valente, 2003). Borgatti et al (2006) and L.Frantz et al (2009) investigated how centrality measures of nodes in networks are affected by the random addition and deletion of nodes and links. Borgatti et al (2006) show that the accuracy of centrality measures declines smoothly and predictably with the amount of error. L.Frantz et al (2009) show that the robustness of centrality measures in a graph is affected by its structural characteristics. However, since unweighted graphs were used for the analyses in these studies, the effects of ignoring link weight and link weight quantization on centrality measures have not been explored. Therefore, we study the effects of link weight quantization on centrality measures through simulations using weighted graphs.

Recently, the use of link weights for social network analysis has become popular, particularly among researchers studying social networks obtained from communication logs (Ehrlich and Cataldo, 2012; Cataldo and Ehrlich, 2012; Gómez et al, 2008). For example, Cataldo and Ehrlich (2012) investigated the performance of software developers through their centrality measures in weighted communication networks. On the other hand, several researchers have proposed to artificially generate link weights in unweighted networks to improve the quality of community detection (Sun, 2014; Berry et al, 2011; Meo et al, 2012; De Meo et al, 2014; Sun et al, 2013). These studies suggest that using link weights improve social network analyses. In contrast, Bonacich (1987) claimed that using link weights may not necessarily lead to better inferences about power and influence. Thus, the amount of benefit we can gain by using weighted networks with continuous values of link weights compared to unweighted graphs or graphs with quantized link weights is still unclear. Therefore, in this study, we aimed to clarify how much benefit can be gained by using a weighted network with continuous link weights compared to an unweighted graph or a graph with quantized link weights.

3 Methodology

We investigate how centrality measures of nodes differ between a weighted undirected graph G, in which link weights have continuous values, and a weighted undirected graph G_n, in which link weights are quantized to take n discrete values, generated from the graph G. Graph G represents the ground truth social network, which the researchers would like to analyze, and graph G_n is a quantized graph, which the researchers use for their social network analyses. By comparing the node rankings between graph G and G_n, we investigate how centrality measures are affected by link weight quantization. If the node ranking does not change between graph G and G_n, the centrality measures are considered to be robust, and in the opposite case, the centrality measures are considered to be not robust.

We randomly generate weighted undirected graphs G using network generation models. Since there are several definitions of links (i.e., social ties among individuals) in social network analyses, the topological structures of the graphs used for the analyses are also different from each other. There are several network generation models for unweighted social networks, but there are not many weighted social network generation models. The following three generation models are popular and we, therefore, use these models to generate graphs with different structural characteristics.

- Community Emergence (CE) model (Kumpula et al, 2009)
  The CE model is a network generation model that models the formation of community structure in a social network. A weighted undirected graph generated by the CE model consists of clusters of nodes, which are densely connected to each other by links with large weights. Between the clusters are a small number of links with small weights.
- Weighted Evolution (WE) model (Barrat et al, 2004)
  The WE model is a network generation model that models the evolution of link weights and the topology created by the existing link weights. A weighted undirected graph generated by the WE model has the feature that the distributions of node degree and link weights follow power laws.
- Weighted Evolving with Community Structure (WECS) model (Li and Chen, 2006)
  The WECS model is a network generation model that models the formation of community structure in a social network by the evolution of link weights and the formation of a topology depending on the existing link weights. A weighted undirected graph generated by WECS model has a cluster structure, a power-law distribution of node degree, and a power-law distribution of link weights.

For comparison purposes, we also use the weighted undirected graph that is an Erdös - Rényi (ER) graph with link weights randomly assigned according to the Pareto distribution. In what follows, we call this model for generating graphs the "Random graph with Random link weight" (RR) model.

There are two types of quantization approaches: linear and non-linear quantization. Since the link weight distributions of social networks are generally skewed, a non-linear approach is expected to be an effective way to reduce quantization errors. In particular, in social networks representing human communication, link weights follow a power-law distribution, and the aforementioned models generate such link weight distributions (Kumpula et al, 2009; Barrat et al, 2004; Li and Chen, 2006). As a representative of the non-linear quantization approach, we use logarithmic quantization, and for comparison purposes, we use linear quantization. In linear quantization we divide the range of link weights into n equal sections, and assign an integer between 1 and n to each section. The integer k associ-

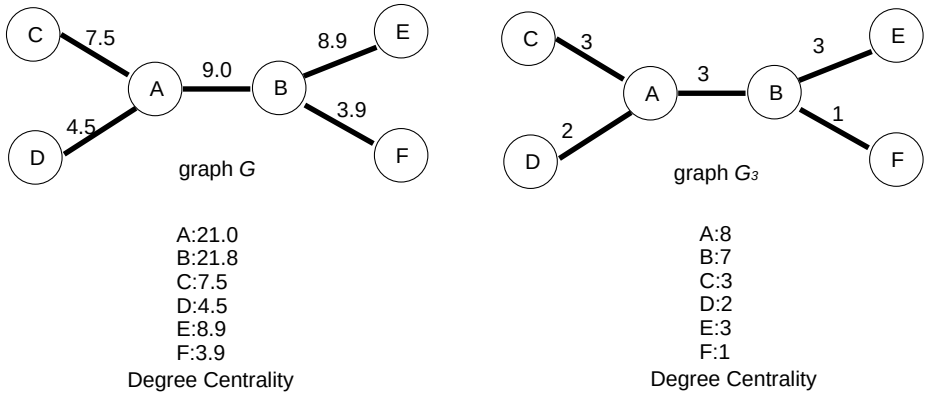

Fig. 2: An example of calculating weighted degree centrality. (The degree centrality of a node is defined as the sum of its link weights.)

ated with a weight w is then given by

$$k = \lceil n \times \frac{w}{w_{max}} \rceil, \tag{1}$$

where w_{max} is the maximum link weight in graph G. In logarithmic quantization we divide the range of link weights into n equal sections after a logarithmic transformation, and assign an integer between 1 and n to each section. The possible link weights in graph G_n are

$$w_{max}^{\frac{k}{n}}, \tag{2}$$

for $1 \leq k \leq n$ and the integer k associated with weight w is given by

$$k = \lceil n \times log_{w_{max}} w \rceil. \tag{3}$$

We consider four centrality measures for weighted networks: degree centrality (Opsahl et al, 2010), betweenness centrality (Opsahl et al, 2010), closeness centrality (Opsahl et al, 2010), and eigenvector centrality (Bonacich, 1972). Degree centrality, betweenness centrality, and closeness centrality are indices that represent the influence of a node on others based on its degree, the proportion of shortest paths between all other node pairs passing through the node, and the shortest path lengths from the node to all other nodes, respectively. Eigenvector centrality is an index that represents the influence of a node on others based on the centrality of adjacent nodes. Figure 2 shows an example of calculating weighted degree centrality in graph G and graph G_3. As shown in this figure, node ranking may change when the link weights are quantized: node B has the highest ranking in terms of degree centrality in the graph G, which has continuous values of link weights, while node A has the highest ranking and node B is in 2nd place in the graph G_3, in which the link weights are quantized to take 3 values.

We rank all the nodes in graphs G and G_n by sorting them in descending order of their centrality measures. Then, we calculate the consistency of node rankings between graphs G and G_n. Following (Borgatti et al, 2006; L.Frantz et al, 2009), we use Top_1, Top_3, $\mathrm{Top}_{10\%}$, $\mathrm{Overlap}_{10\%}$, and R^2as measures of ranking consistency. Top_m is 1 if the most central node in graph G lies in the top m most central nodes in graph G_n, and otherwise it is 0. $\mathrm{Top}_{p\%}$ is

Table 1 Parameter of network generation models (p is the probability to generate edges in the ER model, a and b are the parameters of the Pareto distribution, and other symbols are parameters of the CE, WE, and WECS models defined in (Kumpula et al, 2009; Barrat et al, 2004; Li and Chen, 2006))

CE model		WE model		WECS model		RR model	
δ	1.0	δ	1	M	2	p	0.05
p_δ	8.85×10^{-3}	m	3	m_0	3	a	1.5
p_d	1×10^{-3}			α	0.15	b	1
p_r	1×10^{-3}			β	0.1		
time step	25,000			η	0.1		
				m	2		

1 if the most central node in graph G lies in the top $p\%$ of nodes in graph G_n, and otherwise it is 0. $\mathrm{Overlap}_{p\%}$ is defined as follows. Let A be the set of nodes in the top $p\%$ of graph G, and B be the set of nodes in the top $p\%$ of graph G_n, then $\mathrm{Overlap}_{p\%}$ is defined as $\frac{|A \cap B|}{|A \cup B|}$. $R^2$is the square of the Pearson correlation coefficient between centrality measures in graph G and those in graph G_n. Top_m and $\mathrm{Top}_{p\%}$ are metrics for quantifying the ranking stability of the best (i.e., the most important) node, $\mathrm{Overlap}_{p\%}$ is for highly-ranked nodes, and R^2 is for all nodes. $\mathrm{Overlap}_{p\%}$ and R^2 quantify the stability of broader-range node rankings than Top_m and $\mathrm{Top}_{p\%}$ do, and therefore, we use them as measuers of broad-range node ranking consistency. Using the four network generation models, we randomly generated 2,000 graphs, and calculated the averages and 95% confidence intervals of the various ranking consistency indices. We perform simulation for varying numbers of nodes N and densities ρ, but unless explicitly stated, we use graphs with $N = 100$ and $\rho \simeq 0.05$.

The parameter values used in the network generation models are shown in Tab. 1. Since the 95% confidence intervals were sufficiently small in all cases, only the averages of the ranking consistency indices are shown in the following results.

4 Effects of Link Weight Quantization in Graphs Generated from Network Generation Models

4.1 Effects of Method for Link Weight Quantization and Quantization Level

First, we investigated how the centrality measures are affected by the method for link weight quantization (i.e., linear quantization or logarithmic quantization) and the quantization level.

In what follows, we show the results obtained when using the WECS model as the network generation model and closeness centrality as the centrality measure.

Figure 3 shows the relation between quantization level n and consistency of node ranking (Top_1, Top_3, $\mathrm{Top}_{10\%}$, $\mathrm{Overlap}_{10\%}$, and R^2) when using linear quantization. Note that $n = 1$ is equivalent to ignoring link weight. Figure 4 shows the results for logarithmic quantization.

Comparison of the results for linear quantization (Fig. 3) and for logarithmic quantization (Fig. 4) shows that logarithmic quantization is a more robust method than linear quantization. As we discussed in Section 3, the distribution of link weights of a graph generated from the WECS model follows a power-law distribution. Hence, the quantization levels can be more effectively utilized by using logarithmic quantization rather than linear quantization and, as a result, logarithmic quantization is more robust. More specifically, the quantization error of logarithmic quantization is smaller than that of linear quantization. A mathematical explanation is given in Appendix A. This suggests that, in order to use quantization levels

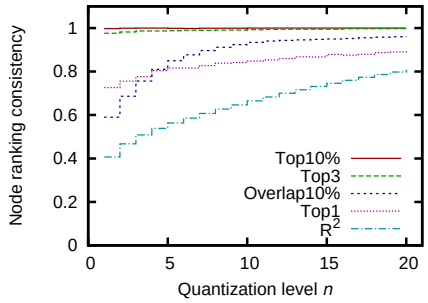

Fig. 3: The relation between the quantization level n and the consistency of node ranking (WECS model, closeness centrality, linear quantization)

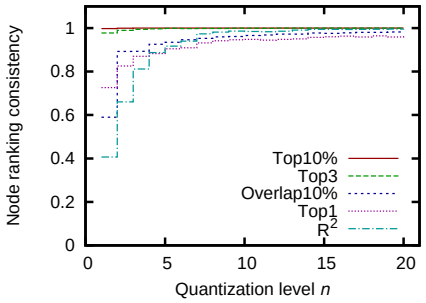

Fig. 4: The relation between the quantization level n and the consistency of node ranking (WECS model, closeness centrality, logarithmic quantization)

effectively, it is important to design questionnaires appropriately to match the distribution of link weights in the graph used in the social network analysis.

Moreover, we can see in Fig. 4 that, while all indices measuring node ranking consistency are more than approximately 0.9 when the quantization level is five or more, Top_1, $\mathrm{Overlap}_{10\%}$, and R^2take small values when the quantization level is less than five. Since the average value of Top_1 is the proportion of times that the most central node in graph G is also the most central node in graph G_n, the average value of Top_1 is naturally smaller than the averages of Top_3 and $\mathrm{Top}_{10\%}$. In contrast, $\mathrm{Overlap}_{10\%}$ and $R^2$are broad-range ranking consistency indices that do not focus only on the most central node.

These results suggest that the effect of link weight quantization is not so significant when the purpose of social network analysis is to infer the most central node. However, these results also suggest that five to eight quantization levels are necessary for determining both the most central node and broad-range node rankings.

4.2 Effects of Graph Size and Density

We next investigated the effects of the graph size and density on the robustness of centrality measures against link weight quantization. In what follows, we consider closeness centrality, the WECS network generation model, and logarithmic quantization.

Figures 5 and Figure 6 show the relation between the quantization level n and $\mathrm{Overlap}_{10\%}$ or Top_1 for graphs of different sizes.

From Fig. 5, we can see that the differences in $\mathrm{Overlap}_{10\%}$ among graphs of different sizes are marginal. On the other hand, Fig. 6 shows that Top_1 decreases as the graph size increases. It is natural to expect that Top_1 for a larger graph takes a lower value than for a smaller graph since it is more difficult to discover the most important node in a larger graph. We, therefore, expect that the effect of graph size on the robustness of centrality measures is not significant. Borgatti *et al.* (Borgatti et al, 2006) report the similar findings that graph size is only weakly related to the robustness of centrality measures against the addition and deletion of nodes and links.

Figure 7 shows the relation between the quantization level n and $\mathrm{Overlap}_{10\%}$ for graphs with different densities in WECS model.

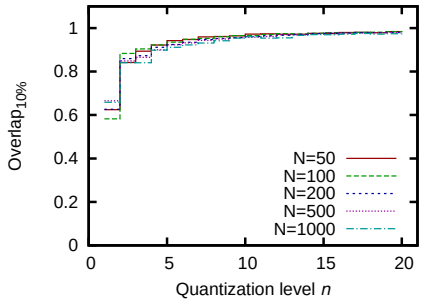

Fig. 5: The relation between the quantization level n and the consistency of the top 10% node ranking, $\mathrm{Overlap}_{10\%}$, for graphs of different sizes (WECS model, closeness centrality, logarithmic quantization).

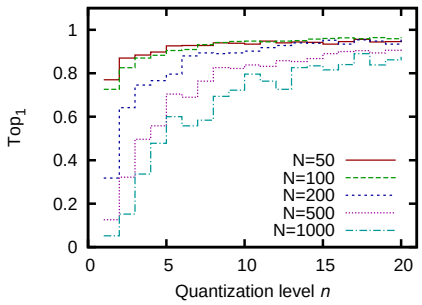

Fig. 6: The relation between the quantization level n and Top_1 for graphs of different sizes (WECS model, closeness centrality, logarithmic quantization).

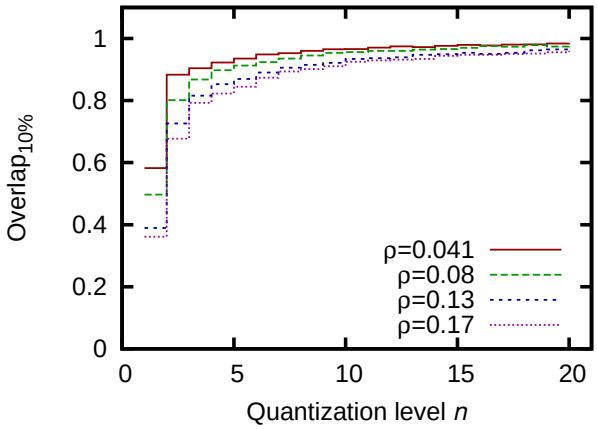

Fig. 7: The relation between the quantization level n and the consistency of the top 10% node ranking, $\mathrm{Overlap}_{10\%}$, for graphs with different densities (WECS model, closeness centrality, logarithmic quantization).

From Fig. 7, we can see that differences in $\mathrm{Overlap}_{10\%}$ among graphs with different densities are marginal when the quantization level n is larger than one. If link weights are ignored (i.e., $n = 1$), the differences in $\mathrm{Overlap}_{10\%}$ among graphs in different densities are significant. However, since most social networks are sparse, it is expected that the effect of graph density on the robustness of centrality measures against link weight quantization is not significant for actual social network analyses. A similar tendency is observed in other models.

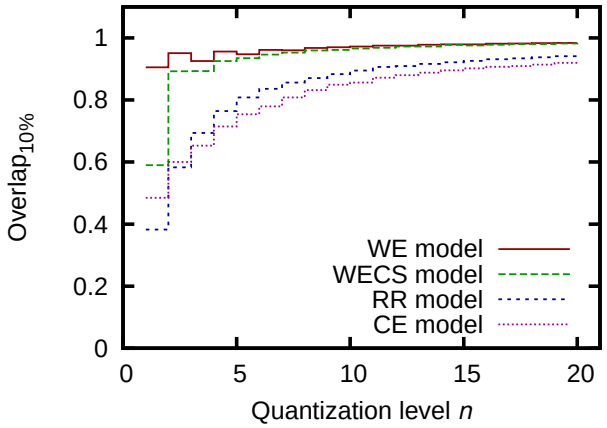

Fig. 8: The relation between the quantization level n and the consistency of the top 10% node ranking, $\mathrm{Overlap}_{10\%}$, for graphs generated by the four network generation models (closeness centrality and logarithmic quantization).

4.3 Effects of Graph Structure

Since there are several definitions of links (i.e., social ties among individuals) in social network analyses, the effect of link weight quantization on centrality measures in graphs with different structural characteristics should be investigated. We, therefore, generated graphs with different structural characteristics using four network generation models, and investigated the relation between the quantization level n and the consistency of node ranking.

In this investigation, closeness centrality was used as the centrality measure and logarithmic quantization was used as the quantization method.

Figure 8 shows the relation between the quantization level n and the consistency of the top 10% node ranking, $\mathrm{Overlap}_{10\%}$, for graphs produced by the four network generation models.

While the curves representing relations between quantization level n and $\mathrm{Overlap}_{10\%}$ are monotonically increasing regardless of the network generation model, their forms are significantly different. In particular, the values of $\mathrm{Overlap}_{10\%}$ when link weights are ignored ($n = 1$) are significantly different for each model. Furthermore, when we focus on quantization levels of two or more steps, we find that the four models can be classified into two categories: models in which the consistency of node rankings does not significantly decrease (WE and WECS models), and models in which the consistency of node rankings decreases rapidly (CE and RR models). This observation suggests that the robustness of centrality measures against link weight quantization is significantly affected by the structural characteristics of graphs rather than the sizes and densities of graphs.

Table 2 shows the averages and standard deviations for various structural characteristics of graphs and several statistics for the link weights calculated from 2,000 graphs randomly generated by the four network generation models. The average and standard deviation of the correlation between node degrees and link weights (i.e., correlation between $w_{i,j}$, which is the weight of link (*i*,*j*), and the sum of the degrees of nodes *i* and *j*) are also shown in Tab. 2. The modularity of a graph with respect to some division of the graph into subgraphs mea-

Table 2 The characteristics of graphs produced by four kinds of network generation models (average μ and standard deviation σ)

	CE model		WE model		WECS model		RR model	
	μ	σ	μ	σ	μ	σ	μ	σ
The characteristics of graphs								
average degree	5.47	0.42	5.88	0.00	4.07	0.06	4.95	0.38
average shortest path length	3.59	0.29	2.39	0.04	3.44	0.14	3.03	0.18
clustering coefficient (Watts, 2003)	0.38	0.03	0.11	0.01	0.10	0.01	0.05	0.01
modularity (Clauset et al, 2004)	0.81	0.03	0.24	0.02	0.51	0.03	0.59	0.04
skewness of the degree distribution	1.16	0.45	3.62	0.39	2.70	0.48	0.38	0.25
kurtosis of the degree distribution	1.97	2.31	14.0	3.57	8.36	3.86	0.00	0.67
Statistics of the link weights								
average	482	68.3	1.98	0.00	3.69	0.46	3.02	1.82
standard deviation	898	190	1.73	0.09	8.69	2.26	8.22	27.2
median	155	27.8	1.37	0.04	1.03	0.16	1.59	0.07
skewness	4.13	1.30	3.47	0.39	5.75	1.29	7.90	3.33
kurtosis	24.8	19.5	14.3	3.67	40.0	19.5	84.1	63.0
maximum of link weights	7618	2988	12.8	1.49	78.3	29.8	110.4	2427.8
correlation coefficient between node degrees and link weights	0.18	0.08	0.57	0.04	0.36	0.07	0.00	0.07

sures how good that division is. The mean and standard deviation of the maximum modularity scores are given from clusterings obtained by using modularity maximization (Clauset et al, 2004).

Figure 8 and Tab. 2 show that if the skewness of the degree distribution is large and the correlation between node degrees and link weights is strong in graph G, then graph G is robust against link weight quantization. Thus, these results suggest that the effect of link weight quantization on centrality measures depends greatly on the characteristics of graphs used in social network analyses. It is intuitive that a strong correlation between node degrees and link weights in a graph results in the robustness of the graph against link weight quantization, since it is expected that the link weights contain little information in such graphs. Note that eigenvector centrality is reported to be robust against random link rewiring in scale-free networks, for which the skewness of the degree distribution is large (Ghoshal and Barabási, 2011). Moreover, the four centrality measures are also known to be robust against the random addition and deletion of nodes and links in scale-free networks (L.Frantz et al, 2009). In a scale-free network, there are nodes with very high degree, and those nodes tend to have very high centrality regardless of whether the graph has weighted links or not.

To confirm that if the skewness of the degree distribution is large and the correlation between node degrees and link weights is strong in a graph, then the graph is robust against link weight quantization, we investigate the relation between the quantization level n and the consistency of node ranking for graphs in which the correlations between node degrees and link weights are weakened. We generate such graphs by swapping link weights randomly in graphs generated by the four network generation models. Figure 9 shows the relation between the quantization level n and $\mathrm{Overlap}_{10\%}$ for graphs generated by the WECS model and for three kinds of graphs in which correlations between node degrees and link weights are weakened by swapping link weights randomly in graphs generated by the WECS model. This result shows that if the correlation between node degrees and link weights is strong in a graph, then the graph is robust against link weight quantization. Figure 10 shows the relation between the quantization level n and $\mathrm{Overlap}_{10\%}$ for graphs in which the correlation coefficient between node degrees and link weights is zero, which were produced by swapping link weights randomly in graphs generated by the four network generation models. This result shows that if the skewness of the degree distribution is large in a graph, then the graph is robust against link weight quantization, even when the correlation coefficient between node degrees and link weights in graphs is zero.

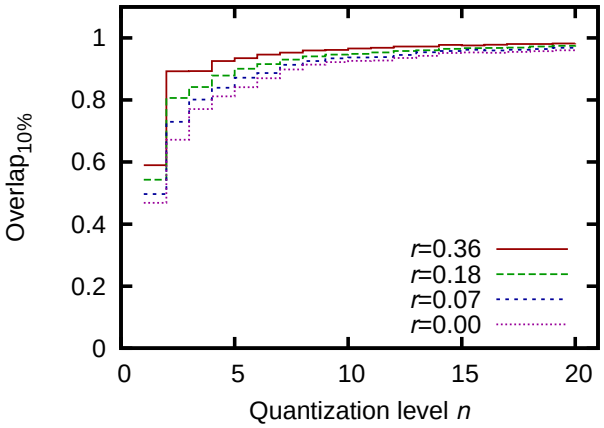

Fig. 9: The relation between the quantization level n and the consistency of the top 10% node ranking, $\mathrm{Overlap}_{10\%}$, for graphs produced by WECS model and three kinds of graphs in which the correlation between node degrees and link weights is weakened by replacing link weights randomly in graphs produced by the WECS model (closeness centrality and logarithmic quantization).

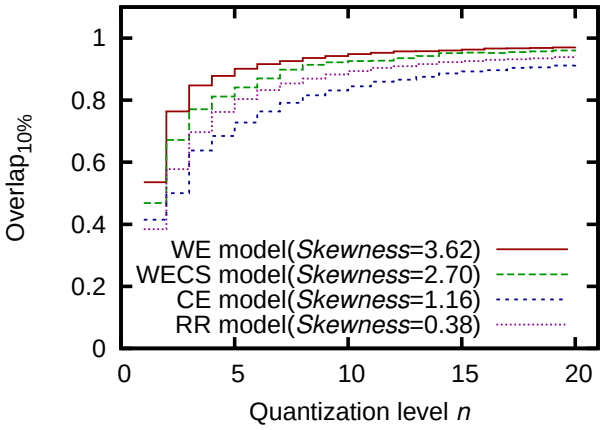

Fig. 10: The relation between the quantization level n and the consistency of the top 10% node ranking, $\mathrm{Overlap}_{10\%,}$ for graphs in which the correlation coefficient between node degrees and link weights is 0, created by swapping link weights randomly in graphs produced by the four network generation models (closeness centrality and logarithmic quantization).

These results suggest that graphs with a highly skewed degree distribution or with a high correlation between node degrees and link weights are robust against link weight quantization.

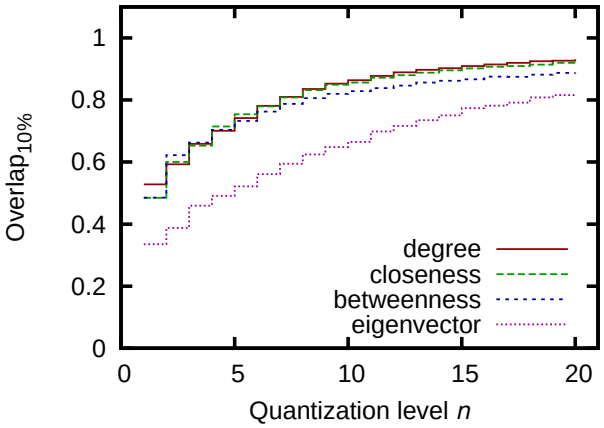

Fig. 11: The relation between the quantization level n and the consistency of node ranking, $\mathrm{Overlap}_{10\%}$, in graphs produced by the CE model (logarithmic quantization)

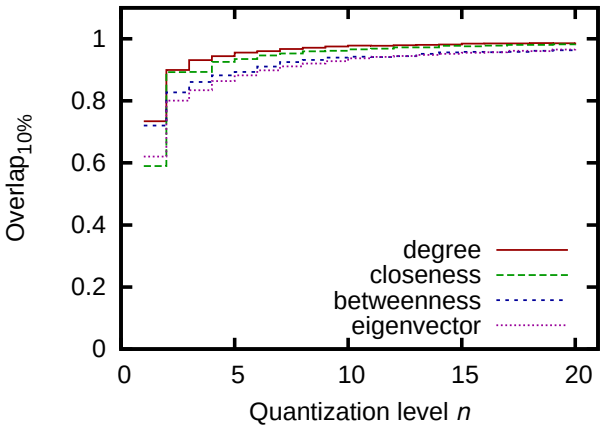

Fig. 12: The relation between the quantization level n and the consistency of node ranking, $\mathrm{Overlap}_{10\%}$, in graphs produced by the WECS model (logarithmic quantization)

4.4 Effects of the Type of Centrality Measure

We next investigated how the effects of link weight quantization differ among the four types of centrality measures. Figures 11 and 12 show the relation between the quantization level n and the consistency of the top 10% node ranking, $\mathrm{Overlap}_{10\%}$, for graphs generated by the CE and WECS models.

Figures 11 and 12 show that the relations between quantization level and node ranking consistency are quite similar for three of the four types of centrality measures. Hence, as we discussed in Section 4.3, these results suggest that the effect of link weight quantization on centrality depends more on the characteristics of the graphs rather than on the type of

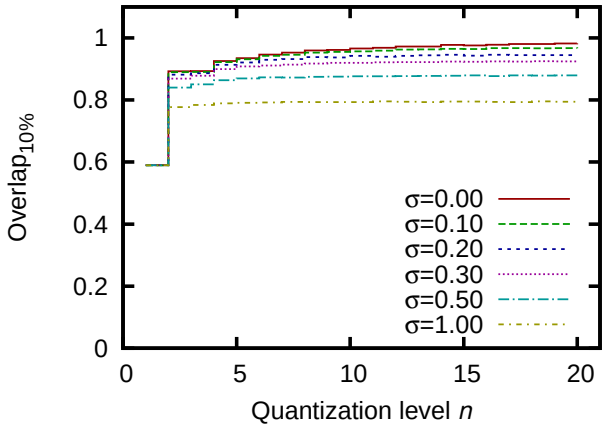

Fig. 13: The relation between the quantization level n and the consistency of the top 10% node ranking, $\mathrm{Overlap}_{10\%}$, for graphs in which link weights contain Gaussian noise with standard deviation σ (WECS model, closeness centrality, and logarithmic quantization).

centrality used in social network analysis. The four types of centrality measures also have a similar robustness against random addition and deletion of nodes and links (Borgatti et al, 2006).

However, Fig. 11 shows that when using eigenvector centrality, the consistency of the node ranking, $\mathrm{Overlap}_{10\%}$, is significantly smaller than for other centrality measures. We note that the robustness of eigenvector centrality for graphs generated by the RR model is found to be similar to that for graphs generated by the CE model. Further investigation is needed to determine the reason why eigenvector centrality is significantly affected by link weight quantization for graphs generated from the CE and RR models.

4.5 Effects of both Link Weight Quantization and Gaussian Noises

We next investigated the effects of both link weight quantization and link weight noise on centrality measures. As we discussed in Section 1, the strengths of social ties are generally known from questionnaires given to the participants in experiments, and it may be difficult for the participants to accurately recognize the strengths of their social ties. Hence, graphs used for social network analysis may contain link weight noise which will affect the centrality measures.

To investigate the effects of link weight noise, we generated a weighted undirected graph G_σ, by adding Gaussian noise to each link weight in graph G. More specifically, for each link weight $w_{i,j}$ in graph G, $w_{i,j}$ is multiplied by η, which is a random variable generated according to the normal distribution $\mathcal{N}(1.0, \sigma^2)$. We then obtain graph G_n by quantizing link weights of graph G_σ in the way explained in Section 3. Note that if $w < 0$, we simply use $k = 1$. In this investigation, we used closeness centrality, the WECS network generation model, and logarithmic quantization.

Figure 13 shows the relation between the quantization level n and $\mathrm{Overlap}_{10\%}$ for graphs in which link weights have Gaussian noise with standard deviation σ. From this

figure, we can see that when the amount of noise is not so large (e.g., $\sigma \leq 0.3$), $\mathrm{Overlap}_{10\%}$ is comparable to the case with no noise (i.e., $\sigma = 0$). The results are similar when we use other network generation models or any other centrality measure except eigenvector centrality. A noise level of $\sigma = 0.3$ means that approximately 70% of link weights are multiplied by 0.7 – 1.3. The actual noise level in social network analysis is unknown, but this result suggests that when the amount of noise is not so large, the effect on the centrality measures can be considered to be not significant. This is because even if Gaussian noise is contained in a link weight, both the link weight with the noise and the original link weight (i.e., the link weight without noise) can take the same value when quantized.

In contrast, we also find that when link weight noise is included, the benefit of increasing the quantization level is small. The increase of $\mathrm{Overlap}_{10\%}$ is small even when we use a much larger quantization level. From this result we suggest that, if the amount of noise is large, using a 2-3 step quantization level is enough, and using a continuous link weight does not have much benefit.

5 Effects of Link Weight Quantization in Real Social Networks

5.1 Overview of Real Social Networks

In this section, we perform experiments with real social networks to validate the results obtained from simulations. We use the following four types of real networks.

- Enron Email Networks (EENs)
  An EEN, which is generated from the Enron Email Dataset (Shetty and Adibi, 2004), represents the frequencies of email communications among 151 employees of the Enron Corporation from April 2000 through March 2002. Nodes in an EEN are the 151 employees. An undirected link (i, j) is created if an email communication, detected from the `To`, `CC`, and `From` fields, exists between employees i and j. The link weight $w_{i,j}$ is defined to be the number of email messages exchanged between employees i and j. Since a large number of communication logs are available from the Enron Email Dataset (252,759 emails), we generated an undirected graph from each monthly email log.
- Online Community Network (OCN)
  The OCN, which is generated from log data of messages exchanged in an online community in which participants are students in the University of California, Irvine (Opsahl and Panzarasa, 2009), represents the frequencies of communication by direct messaging among 1,899 students. Nodes in the OCN are the 1,899 students. An undirected link (i, j) is created if a direct message communication exists between students i and j. The link weight $w_{i,j}$ is defined to be the number of direct messages exchanged between user i and j. An undirected graph is generated from all 59,835 messages.
- Forum Network (FN)
  The FN, which is generated from log data of messages exchanged in online bulletin boards in which participants are students in the University of California, Irvine (Opsahl, 2013), represents frequencies of communication in bulletin boards among 899 students. Nodes in the FN are the 899 students. An undirected link (i, j) is created if a reply communication exists between students i and j. The link weight $w_{i,j}$ is defined to be the number of replies in the bulletin boards between users i and j. An undirected graph is generated from all messages.
- US-AN (US Airport Network)

Table 3 The characteristics of four real networks (average μ and standard deviation σ are given for EEN networks).

	EEN		OCN	FN	US-AN
	μ	σ			
The characteristics of graphs					
number of nodes	94.5	28.4	1899	899	500
average degree	4.33	1.45	14.5	159	11.92
average shortest path length	3.76	0.58	3.05	1.88	2.99
clustering coefficient (Watts, 2003)	0.31	0.10	0.06	0.50	0.35
modularity (Clauset et al, 2004)	0.66	0.10	0.77	0.15	0.28
skewness of the degree distribution	2.16	1.33	4.24	0.78	3.38
kurtosis of the degree distribution	8.54	12.8	25.6	-0.03	12.3
Statistics of the link weights					
average	4.63	0.99	4.32	15.6	304640
standard deviation	8.05	2.21	7.95	30.6	440586
median	1.96	0.46	2	7	141045
skewness	4.94	2.30	8.18	10.37	3.18
kurtosis	36.9	36.0	107	232	14.3
maximum of link weights	69.5	32.1	184	1569	4507980
correlation coefficient between node degree and link weights	0.10	0.17	0.07	0.39	0.49

The US-AN, which is generated from the dataset used in (Colizza et al, 2007), represents relationships among the 500 busiest commercial airports in the United States. Nodes in the US-AN are the 500 airports. An undirected link $(i,\ j)$ is created if a flight is scheduled between airports i and j. The link weight $w_{i,j}$ is defined to be the number of seats available on the scheduled flights between airports i and j. Since the US-AN is not a social network, we use this network for comparison purposes.

We use these four types of networks as graph G. We use the quantization methods and measures of ranking consistency described in Section 3. Table 3 shows the averages of several indices for the structural characteristics of the four types of real networks, several statistics for the link weights, and correlations between node degree and link weights. Since we obtain multiple EEN graphs, the standard deviations and the averages of several statistics are shown.

5.2 Effects of Quantization Level

Figure 14 shows the relation between quantization level n and consistency of node ranking (Top_1, Top_3, $\mathrm{Top}_{10\%}$, $\mathrm{Overlap}_{10\%}$, and R^2) when using closeness centrality and logarithmic quantization. Since multiple graphs of EENs are available, we use EENs as the real social networks.

We can see from Fig. 14 that while all indices measuring node ranking consistency are approximately more than 0.9 when the quantization level is ten or larger, all indices except $\mathrm{Top}_{10\%}$ take small values when the quantization level is less than five in an EEN. This result suggests that the effect of link weight quantization is so significant in an EEN that 5 to 10 quantization levels are necessary.

This tendency is similar to the result from WECS model (Fig. 4). However, this result also suggests that an EEN is less robust than a graph generated by the WECS model. One possible explanation is given by the skewness of the degree distribution and the correlation between node degree and link weights in EENs and graphs generated by WECS model. Both the skewness of degree distribution and the correlation between node degree and link weights in an EEN are smaller than those of graphs generated by the WECS model (see Tabs. 2 and 3).

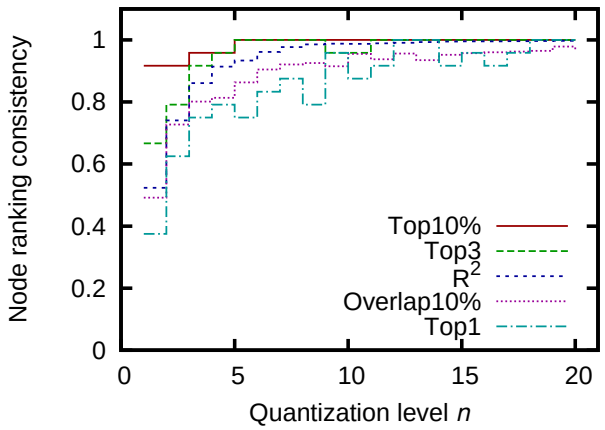

Fig. 14: The relation between the quantization level n and the consistency of node ranking in EEN (closeness centrality and logarithmic quantization)

5.3 Effects of Graph Structure

We finally compare the four types of real networks in order to investigate the effects of graph structure on the robustness of centrality measures. As shown in Tab. 3, the sizes and densities are significantly different among the four types of networks. However, we have shown that the effects of size and density on the robustness of centrality measures are marginal. We expect that the effects of graph structure can be investigated through comparison of the four types of networks.

Figure 15 shows the relation between the quantization level n and $\mathrm{Overlap}_{10\%}$ for four types of real networks.

Figure 15 shows that while the differences in $\mathrm{Overlap}_{10\%}$ among the four types of networks becomes small as the quantization level n increases, the difference is quite large when the quantization level n is small. In particular, when the quantization level n is 1, $\mathrm{Overlap}_{10\%}$ for the US-AN is more than 0.8 while $\mathrm{Overlap}_{10\%}$ for the other networks is approximately 0.5 – 0.6.

This result again suggests that the effect of link weight quantization on centrality measures depends on the characteristics of the graphs. From Tab. 3, we can see that both the skewness of degree distribution and the correlation between node degrees and link weights in the US-AN are high. Hence, it is expected that graphs with highly skewed degree distributions or with a high correlation between node degrees and link weights are robust against link weight quantization.

6 Conclusion and Future Work

The main conclusions of this work can be summarized as follows.

– We found that the effects of link weight quantization on the centrality measures are not significant when determining the most important node. Top_3 and $\mathrm{Top}_{10\%}$, which focus on the accuracy of determining the most important node, are high (above 0.9)

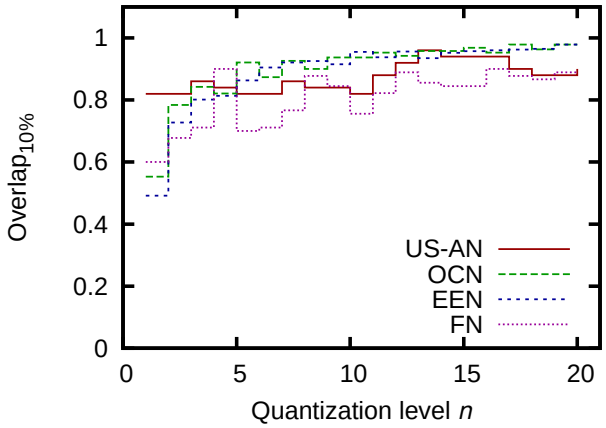

Fig. 15: The relation between the quantization level n and the consistency of the top 10% node ranking, $\mathrm{Overlap}_{10\%}$ in four types of real networks (closeness centrality and logarithmic quantization)

in most cases. This suggests that in order to find the most important node, the benefit of using weighted links is small. For such objectives, questionnaire-based, traditional social network analysis should be enough.

– In contrast, we also found that a 5–8 quantization level is needed to determine not only the most important node but also other important nodes. $\mathrm{Overlap}_{10\%}$, which measures broad-range node ranking consistency, takes a small value when the quantization level is small. Therefore, a relatively higher quantization level of 5–8 steps is necessary to obtain accurate broad-range node ranking. As another way of accurately obtaining broad-range node ranking, we suggest the use of continuous link weights for social network analyses that need broad node rankings. It is possible to obtain continuous values for link weights by analyzing log data for communications, such as phone calls, emails, and message exchanges in social media. Such data should be useful for social network analyses.
– Another key finding in this paper is that the robustness of centrality measure is heavily dependent on the structure of the network. This suggests that when analyzing networks with highly skewed degree distributions, like the WE model, we need not consider social tie strength. But in most synthetic and real networks in our experiments, the effect of link weight quantization is not negligible.

One limitation of this paper is that we only focus on link weight quantization. In more realistic situations, several types of errors are present in a network. Hence, we are planning to investigate the robustness of centrality measures in the presence of heterogeneous errors (e.g., node and link addition or deletion). This research is a first step toward understanding the effects of errors in link weights. More research is, therefore, needed. For instance, further investigation of the effects of link weight noise is one important future task.

Another limitation is investigation of the effects of network structure. Characteristics such as core-periphery distinctions and hierarchies are important characteristics observed in many types of social networks (Cattani and Ferriani, 2008; Uzzi and Spiro, 2005; Cataldo and Ehrlich, 2012). However, there are not many models that generate weighted networks, so we used popular weighted network generation models. As a result, topologies such as

core-periphery distinctions and hierarchies were not considered in our experiments. Investigating the effects of such structural characteristics on the robustness of centrality measures is important future work.

One way to obtain several types of topologies, such as core-periphery distinctions and hierarchies, is to generate unweighted networks, and apply a weighting procedure to generate artificial link weights for the network (Sun, 2014). However, this approach is new, and the relation between artificially generated weights and social ties in the real world has not yet been explored. But, as the relation becomes clearer, we expect that this method will be used in our experiments. Another way to investigate the effects of network structure is using various real social networks. Public datasets are available for instance in the the Stanford Network Analysis Project (SNAP) (Leskovec and Krevl, 2014). Such datasets should be useful to investigate the effects of network characteristics, and confirm our results.

Acknowledgements

The authors would like to thank Dr. Makoto Imase for valuable discussions, and anonymous reviewers for their insightful comments.

References

Barrat A, Barthélemy M, Vespignani A (2004) Modeling the evolution of weighted networks. Physical Review E 70(6):66,149

Batallas D, Yassine A (2006) Information leaders in product development organizational networks: Social network analysis of the design structure matrix. IEEE Transactions on Engineering Management 53(4):570–582

Berry JW, Hendrickson B, LaViolette RA, Phillips CA (2011) Tolerating the community detection resolution limit with edge weighting. Physical Review E 83(5):056,119

Bonacich P (1972) Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology 2(1):113–120

Bonacich P (1987) Power and centrality: A family of measures. American journal of sociology 92(5):1170–1182

Borgatti SP (2006) Identifying sets of key players in a social network. Computational & Mathematical Organization Theory 12(1):21–34

Borgatti SP, Carley KM, Krackhardt D (2006) On the robustness of centrality measures under conditions of imperfect data. Social Networks 28(2):124–136

Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. Science 323(5916):892–895

Cataldo M, Ehrlich K (2012) The impact of communication structure on new product development outcomes. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12), pp 3081–3090

Cattani G, Ferriani S (2008) A core/periphery perspective on individual creative performance: Social networks and cinematic achievements in the Hollywood film industry. Organization Science 19(6):824–844

Clauset A, Newman M, Moore C (2004) Finding community structure in very large networks. Physical Review E 70(6):066,111

Colizza V, Pastor-Satorras R, Vespignani A (2007) Reaction–diffusion processes and metapopulation models in heterogeneous networks. Nature Physics 3(4):276–282

Costenbader E, Valente T (2003) The stability of centrality measures when networks are sampled. Social Networks 25(4):283–307

Creswick N, Westbrook J (2010) Social network analysis of medication advice-seeking interactions among staff in an Australian hospital. International Journal of Medical Informatics 79(6):116–125

Cross R, Parker A (2004) The hidden power of social networks: Understanding how work really gets done in organizations. Harvard Business Press

De Meo P, Ferrara E, Fiumara G, Provetti A (2014) Mixing local and global information for community detection in large networks. Journal of Computer and System Sciences 80(1):72–87

Ehrlich K, Cataldo M (2012) All-for-one and one-for-all?: A multi-level analysis of communication patterns and individual performance in geographically distributed software development. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12), pp 945–954

Freeman L (1979) Centrality in social networks conceptual clarification. Social networks 1(3):215–239

Ghoshal G, Barabási A (2011) Ranking stability and super-stable nodes in complex networks. Nature Communications 2(394):1–7

Gómez V, Kaltenbrunner A, López V (2008) Statistical analysis of the social network and discussion threads in slashdot. In: Proceeding of the 17th International Conference on World Wide Web (WWW '08), pp 645–654

Granovetter MS (1973) The strength of weak ties. American journal of sociology 78(6):1360–1380

Kamei Y, Matsumoto S, Maeshima H, Onishi Y, Ohira M, Matsumoto K (2008) Analysis of coordination between developers and users in the apache community. Open Source Development, Communities and Quality 275:81–92

Kim P, Jeong H (2007) Reliability of rank order in sampled networks. The European Physical Journal B-Condensed Matter and Complex Systems 55(1):109–114

Kumpula J, Onnela J, Saramäki J, Kertész J, Kaski K (2009) Model of community emergence in weighted social networks. Computer Physics Communications 180(4):517–522

Lee SH, Kim PJ, Jeong H (2006) Statistical properties of sampled networks. Physical Review 73(1):016,102

Leskovec J, Krevl A (2014) SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`

LFrantz T, Cataldo M, Carley K (2009) Robustness of centrality measures under uncertainty: Examining the role of network topology. Computational and Mathematical Organization Theory 15:303–328

Li C, Chen G (2006) Modelling of weighted evolving networks with community structures. Physica A: Statistical Mechanics and its Applications 370(2):869–876

Meo PD, Ferrara E, Fiumara G, Provetti A (2012) Enhancing community detection using a network weighting strategy. Information Sciences 222(10)

Opsahl T (2013) Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. Social Networks 35(2):159 – 167

Opsahl T, Panzarasa P (2009) Clustering in weighted networks. Social Networks 31(2):155–163

Opsahl T, Agneessens F, Skvoretz J (2010) Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks 32(3):245–251

Shetty J, Adibi J (2004) The Enron email dataset database schema and brief statistical report. Tech. rep., Information Sciences Institute, University of Southern California

Sun PG (2014) Weighting links based on edge centrality for community detection. Physica A: Statistical Mechanics and its Applications 394:346 – 357

Sun PG, Gao L, Yang Y (2013) Maximizing modularity intensity for community partition and evolution. Information Sciences 236:83–92

Tsugawa S, Ohsaki H, Imase M (2012) Inferring leadership of online development community using topological structure of its social network. Journal of the Infosocionomics Society 7(1):17–27

Uzzi B, Spiro J (2005) Collaboration and creativity: The small world problem. American journal of sociology 111(2):447–504

Valente T, Watkins S, Jato M, Straten AVD, Tsitsol L (1997) Social network associations with contraceptive use among Cameroonian women in voluntary associations. Social Science & Medicine 45(5):677–687

Watts DJ (2003) Small worlds: the dynamics of networks between order and randomness. Princeton Univ Pr

Watts DJ (2007) A twenty-first century science. Nature 445(7127):489

Appendix

A Quantization Errors of Linear and Logarithmic Quantization

We derive the quantization errors of linear and logarithmic quantization. Without loss of generality, we assume link weights are distributed over $[1, m]$. As a representative skewed distribution, we assume link weights follow a truncated Pareto distribution with probability density function defined as following equation.

$$p(x) = \begin{cases} \frac{\alpha x^{-\alpha-1}}{1-m^{-\alpha}} & (1 \leq x \leq m) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Let n be the quantization level. Then, the expected value of the error of linear quantization is given by

$$\sum_{k=1}^{n} \int_{1+\frac{(m-1)}{n}(k-1)}^{1+\frac{(m-1)}{n}k} p(x)(1 + \frac{m-1}{n}k - x)dx, \tag{5}$$

and the expected value of the error of logarithmic quantization is given by

$$\sum_{k=1}^{n} \int_{m^{\frac{k-1}{n}}}^{m^{\frac{k}{n}}} p(x)(m^{\frac{k}{n}} - x)dx. \tag{6}$$

The relation between quantization level n and quantization error is shown in Fig. 16. The relation between the parameter α and quantization errors is shown in Fig. 17.

These figures show that when the link weight distribution is the truncated Pareto distribution, which is a skewed distribution, quantization error of logarithmic quantization is smaller than that of linear quantization.

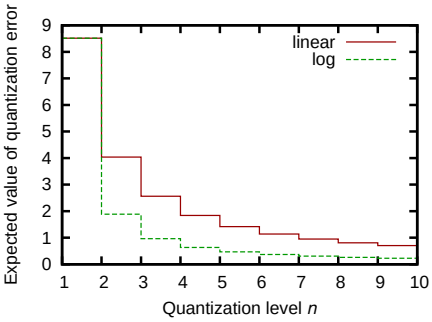

Fig. 16: Relation between quantization level n and expected value of quantization error (parameter $\alpha = 3$, maximum link weight $m = 10$).

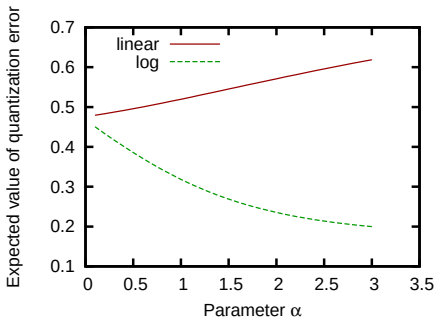

Fig. 17: Relation between parameter α and expected value of quantization error (quantization level $n = 10$, maximum link weight $m = 10$).